Communications Engineering Branch
Annual Report 2002
Submitted September 2002
George R. Thoma

The Communications Engineering Branch is engaged in applied research and development in image engineering and communications engineering motivated by NLM's mission-critical tasks such as document delivery, archiving, automated production of MEDLINE records, Internet access to biomedical multimedia databases, and imaging applications in support of medical educational packages employing digitized radiographic, anatomic, and other imagery. In addition to applied research, the Branch also developed and maintains operational systems for production of bibliographic records for NLM=s flagship database, MEDLINE.

Research areas include: content-based image indexing and retrieval of biomedical images, document image analysis and understanding, image compression, image enhancement, image feature identification and extraction, image segmentation, image retrieval by *query by image content*, image transmission and video conferencing over networks implemented via asynchronous transfer mode (ATM) and satellite technologies, optical character recognition (OCR) and man-machine interface design applied to automated data entry.

CEB also maintains archives of large numbers of digitized spine x-rays and bit-mapped document images that are used for intramural and outside research purposes.

Information on these projects appears at *http://archive.nlm.nih.gov/*

Document imaging for the biomedical end-user

The focus of this research area is to apply document image processing and digital imaging techniques to document delivery via the Internet. It addresses NLM's mission of providing document delivery to end users and libraries. The three active projects in this area are DocView, DocMorph and MyMorph.

DocView

First released in January 1998, this Windows-based client software, subsequently improved over several generations, has over 12,000 users in 175 countries (a 20% increase over 2001). DocView facilitates the delivery of library documents directly to the patron via the Internet in multiple ways, but it is most commonly used by library patrons to receive scanned journal articles from libraries that use Ariel software for interlibrary loan services. While Ariel®, a product of the Research Libraries Group, is used by libraries and document suppliers routinely to send documents via Internet to similar organizations, there are few options for *end users* to directly receive them.

Once documents in bitmapped image form are received, the user may use DocView to retain them in electronic form, view the images, organize them into "folders" and "file cabinets", electronically bookmark selected pages, manipulate the images (zoom, pan, scroll), copy and paste images, and print them if desired. DocView also serves as a TIFF viewer for compressed images received through the Internet by other means, such as web browsers. Users may receive document images either via Ariel FTP or Multipurpose Internet Mail Extensions (MIME) protocols. With DocView, users may also *forward* documents to colleagues for collaborative work.

DocMorph

The function of the DocMorph server is to enable online users to convert files from one format to another for easier exchange or delivery. It therefore serves as an important resource for librarians. Of the more than 6,000 registered users, twice the number last year, many are biomedical document delivery librarians.

DocMorph allows the conversion of more than 50 different file formats to PDF, for instance, to enable multi-platform delivery of documents. Also, by combining OCR with speech synthesis, DocMorph enables the visually impaired to use library information. It has been used by librarians for the blind and physically handicapped to convert documents to synthetic speech recorded onto audio tapes for blind patrons. Applications such as these are included in the more than 77,000 jobs submitted to DocMorph to date, representing 800,000 pages or 75 GB of data.

The file types handled by DocMorph include: all file types from RLG's Ariel system, primarily for users receiving Ariel documents from their libraries; 9 TIFF types including uncompressed, G3, G4, monochrome and color; word processing including Word, Word for Mac, WordPerfect; Excel (.XLS) and PowerPoint (.PPT) files; JBIG and JPEG compressed files; DICOM; BMP; PSD, Adobe Photoshop; PCD, PhotoCD produced by digital cameras; and others.

The tools and subsystems used to achieve this include: inhouse imaging software written in C, C++ and assembly language; Kodak Image Edit control (available in the Windows operating system); MS Word to handle word processing file types; Image Magick, a freely available package; GhostScript, an Open Source package used with PostScript and PDF files.

While DocMorph's OCR facility converts TIFF images to text (and then to speech), there is no equivalent functionality for PDF and PostScript files. To extend DocMorph for this purpose, we integrated Ghostscript into DocMorph to implement a two step process: first converting PDF and PostScript files to TIFF by Ghostscript, and then have the OCR system convert the TIFF files to text and synthesized speech.

MyMorph

Research on DocMorph usage and the availability of new technology have pointed out new opportunities for improvements. The recent availability of Simple Object Access Protocol (SOAP) that combines XML with HTTP has allowed us to create a web service that significantly improves the DocMorph function used 75 percent of the time, viz., the conversion of files to PDF. This web service (MyMorph) consists of a Windows-based client software and modifications to DocMorph for accommodating SOAP.

Inhouse testing has shown that MyMorph significantly improves user productivity compared to the (conventional) use of DocMorph through a web browser, particularly for users who need to convert large numbers of files to PDF. This is accomplished by reducing the time required for users to interact with the software. Test results show that MyMorph reduces file conversion time from hours to seconds for users with slow modem-based Internet access. The software was released for beta testing in June 2002, and in three months more than 400 people had registered and had begun using MyMorph. The system provides a built-in survey that is presented to users thirty days after installation. This user survey is expected to provide feedback to identify software bugs as well as point to needed improvements in both MyMorph and DocMorph.

The MyMorph SOAP web service promises to position DocMorph as a building block facility serving as an information processing resource. For example, it will facilitate the integration of DocMorph in other systems and future projects such as bulk file migration (useful for the preservation of electronic resources) and language translation. The design of MyMorph has been documented in a paper published in the proceedings of the InfoToday 2002 conference.

Document image analysis and understanding

Research into DIAU combined with database design, GUI design for workstations, image processing, string pattern matching, lexical analysis, speech recognition and related areas underlie the development of MARS (*Medical Article Records System)*, a system to automate the production of MEDLINE records from biomedical journals. From bitmapped images of the first page of the articles, this system is designed to automatically extract the article title, author names, affiliations and the abstract. DIAU research centers on the identification of rules for algorithms for page segmentation, zone labeling, OCR error correction, affiliation ranking and other steps. Operators enter fields (other than the ones automatically extracted), as well as perform text verification before the records are made available to indexers. Ongoing research in this area is described below:

Affiliations correction. While other fields are detected with a high degree of accuracy, correct detection of the affiliations field remains problematic. The reasons for this include: words in italics and very small font size, as well as the fact that indexing conventions require the inclusion of only the first author's affiliation and the removal of all others.

In 2002, we pursued an approach that involved exploiting author names and zip codes (which are usually recognized correctly) as cues to detecting the affiliations. To support this investigation, two tools were developed: a test data generator (TDG) and a test program. The TDG generates input data for the test program consisting of the OCR output and correct text from the MARS production database. The TDG was used to extract author and affiliation data from 1,228 articles processed by MARS during December 2001. Of the 1,228 author names tested, almost 50% were found in the existing author/affiliation database. The test program builds on a previously developed C++ class that generates a similarity score between an affiliation in the OCR output ("OCR affiliation") and a correct one from the database. The similarity score is a function of the Levenshtein edit distance, i.e., the number of changes, additions and deletions necessary to match one affiliation with another. This test program allows our research assistants to view, for a given author name or zip code, three items: the OCR affiliation, an affiliation already in the database and the correct (verified) affiliation. For each such triplet, the research assistant selects from the OCR affiliation or the database affiliation the one more like the correct affiliation, and this choice and the score are recorded. The objective is to determine a similarity score threshold that can be used to automatically predict (with confidence) that the affiliation found in the database is close to that intended by the potentially noisy OCR affiliation. Initial testing shows an improvement in matching that increased the true positive rate (106 to 229) and decreased the false positive rate (45 to 35).

In order to automate this process in the MARS production system, we are developing a daemon program called FindAffiliation. Its design, based on the code in the test program mentioned above, is modular so that as we improve the scoring algorithm, the improvements may be readily incorporated. In the production MARS system, FindAffiliation will run after the Reformat process.

Greek character recognition. While the 5-engine OCR system used in MARS has shown a high degree of accuracy and reliability, it does not recognize Greek letters and biomedical symbols. To address this problem, research was conducted toward a prototype recognition system based on features calculated from the output of *multiple* OCR systems, string pattern matching, and a set of rules derived from an analysis of document content, journal specifications, and medical dictionaries. Our technique uses two passes of a document image page through OCR systems designed for different languages.

The process consists of six steps: (1) scan journal pages, (2) perform the first pass using our 5-engine Prime Recognition OCR system that is limited to English, (3) identify low-confidence words (words that contain one or more characters recognized with low confidence by the Prime system), (4) perform the second pass on these low-confidence characters and words using two multilingual OCR systems, FineReader and Recognita, that combine English and Greek, (5) apply string pattern matching between these low-confidence words and similar words obtained from previous steps, and (6) finally, apply a set of rules derived from document-specific information and medical dictionaries to recognize these low-confidence words.

Document content information and journal specifications are derived from an analysis of the page contents of each journal. The low-confidence words containing Greek characters from previous documents are analyzed and their features (contents, attributes, and frequencies of occurrence) are recorded for use in recognizing characters in subsequent documents. Preliminary evaluation conducted on a sample of medical journal page images shows that the system is capable of improving the recognition of Greek characters embedded within predominantly English language text: 89% of the Greek characters were correctly identified. A paper describing this research was published in the Proceedings of the 6th World Multiconference on Systems, Cybernetics and Informatics in July 2002.

Author name problem. In 2002, the format of author names in MEDLINE was changed. This transition required modifications in our Reformat software to accommodate the new format. The old method formatted John A. Smith as *Smith JA*; the new method formats it as *Smith John A*, where Smith is the ‘lastname’ and John A is designated as ‘othername’. We programmed our Reformat software with the latest formatting rules, and modified our Reconcile software to enable the operator to confirm which is the lastname and which the othername, using a library from the Reformat module to do a “best guess”. Using this library the Reconcile software was modified to display two columns showing the lastname and othername, which is confirmed by the operator.

Automated rule extraction. The MARS system relies on image analysis and lexical analysis algorithms to correctly extract bibliographic data from images. These algorithms are based on rules constructed from features extracted from the layout geometry and OCR output. To date, 3,058 journal titles have been tested, and 2,376 titles can be processed by MARS, i.e., there are suitable rules for these. Of these, bibliographic data is supplied directly by publishers for 853, leaving 1,523 titles as MARS-compliant. Since there are 4500+ titles indexed at NLM, this still leaves 1,500 titles for which no rules have been extracted. Our goal is to develop a method to rapidly extract features from which rules may be constructed for the zoning and labeling algorithms.

Research was conducted toward developing a technique for extracting rules by combining the output of the ZoneCzar module that segments and labels contiguous text regions with the corrected output of the Reconcile workstation. Two versions of the Wagner-Fischer string-matching algorithm (both character- and word-based) were developed to correlate the possibly incorrect text from ZoneCzar and the corresponding text corrected by the Reconcile operator. Using Visual C++, we developed a prototype tool (ZoneMatch) that reads the contents of tables in the MARS database, matches the text in the labeled zones (for author, title, affiliation, abstract) using the string-matching algorithm, and displays the outputs of ZoneCzar and ZoneMatch for visual comparison.

On a test set of 304 page images from indexed journals, we found average matching accuracy (ranging from 96.71% for authors to 99.67% for title and affiliation), and the average time taken to match (0.06 s for title to 3.29 s for abstract). This research will proceed toward a practical implementation of a module suitable for the MARS production system.

Ground Truth data. We believe that the publication of ground truth data from the large set of images and extracted data collected in MARS would be an important contribution to the field by facilitating the development of new document analysis methodologies. Accordingly, the effort to develop the mechanism to disseminate this ground truth data for research by the computer science and informatics communities is under way. The tasks accomplished are as follows:

First, a program was developed to export the MARS production data into XML format, using Visual C++ and Xerces, part of the Apache XML Project (http://xml.apache.org). This export program allows an end user to use a wizard style approach to pick destination directories for the XML/TIFF files, the journal titles and the specific page images desired. The application then converts the MARS data to XML data. In addition, the export routines create line and word-level information that, in some cases, might not have been part of the MARS production data, but would be useful for future research.

The second task was to select the initial ground truth data set. We have identified approximately one thousand page images from different journals to accommodate a broad range of page layout types for this purpose.

The third task was to design and develop a website to serve as a repository for the ground truth data and tools to analyze the data. The infrastructure of the website was developed using IIS 5, UltraDev 4.0, Flash 5.0, and SQL Server 2000, and is completed. As images and XML data are verified for accuracy, they will be placed on the website.

We are now engaged in the following steps:

We are using the freely available TrueViz (from the University of Maryland) to visually inspect and correct data at the zone level, such as accurate segmentation of the page into regions, and accurate labeling of these regions as article title, author, affiliation, and abstract, the bibliographic fields of greatest interest.

Using the TrueViz software as a starting point, we have begun the design of an analysis tool for use by researchers. This tool would be a web application launchable from a web browser, and hence easier to use than TrueViz that requires a user to download and install it. Also, the tool will be designed to allow the entry and display of two datasets so that researchers may compare the output of their algorithms against the ground truth data. In addition, our tool will be designed to support batch processing so that multiple algorithms can use the ground truth data concurrently.

As a final step in the development, we intend to provide a user registration facility and a community web board for comments and discussion among researchers.

Accommodating non-compliant journals. Since we want MARS to accommodate all journals indexed in MEDLINE, we decided to process non-compliant ones (for which rules as yet have not been extracted and algorithms developed) as they come into the production area, and made modifications to the Edit workstation to allow the operator to correct the labeling which could, in general, be incorrect. The Edit operator would view the page image with zones and labels as identified by MARS, and then be able to correct them if necessary. Also, as an indication of OCR accuracy, each labeled zone would post a 'figure of merit' (percentage of low confidence characters in the zone) to allow the operator to call for a rescan if necessary. In addition, a copy/paste function is added to allow the operator to transfer correct text from one field in which it incorrectly appears to the right field. Our hypothesis is that overall system performance improves when these corrections are done early in the workflow instead of waiting till the final Reconcile stage.

Biomedical imaging and multimedia database R&D

The goal of this program is to address fundamental questions that arise in the handling, organization, storage, access and transmission of very large electronic files in general and digitized x-rays in particular. A special focus is research into these topics as applied to heterogeneous multimedia databases consisting of both images and text. This work has evolved from a previous project named DXPNET for *Digital Xray Prototype Network* conducted in collaboration with two other agencies, the National Center for Health Statistics (NCHS) and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS).

Biomedical image-related work in CEB this year consisted of (1) enhancements to the WebMIRS system; (2) enhancements to the Digital Atlas of the Spine; (3) continued support of an on-line, publicly accessible archive of 17,000 digitized x-ray spine images; (4) reasearch into Content Based Image Retrieval (CBIR), and (5) creation of an Image Processing Resources Page on the CEB Web site.

WebMIRS. The Web-based Medical Information Retrieval System is a Java application that allows remote users to access data from two surveys conducted by the National Center for Health Statistics. These are the National Heath and Nutrition Examination Surveys II and III (NHANES II and III), carried out during the years 1976-1980 and 1988-1994, respectively. The NHANES II database accessible through WebMIRS contains records for about 20,000 individuals, with about 2,000 fields per record; the NHANES III database contains records for about 30,000 individuals, with more than 3,000 fields per record. In addition, the 17,000 x-ray images collected in NHANES II may also be accessed with WebMIRS and displayed in low-resolution form. WebMIRS allows a user to control a graphical user interface to construct a query for the NHANES II or NHANES III data. A sample query might be equivalent to the English statements: "Find records for all individuals who reported chronic back pain. Return their age, sex, race, age when the pain began, and longest duration of pain. Also, return the record data required for statistical analysis and display their x-ray images." WebMIRS allows the user to save the returned data to the local disk drive, where it may be analyzed with appropriate statistical tools such as the commercially available SAS and SUDAAN software.

The WebMIRS NHANES II database also contains vertebral boundary data that was collected by a board-certified radiologist for 550 of the 17,000 x-ray images in WebMIRS. This data consists of (x,y) coordinates for approximately 20,000 points on the vertebral boundaries in the cervical and lumbar spine images. Users may do queries that use both radiological and/or health survey data. An example of this type of query is: "Find records for all persons having low back pain (health survey data) *and* fused lumbar vertebrae (radiological data)". The boundary data points are displayable on the WebMIRS image results screen and may be saved to the user's local disk.

WebMIRS was used in two semesters of a graduate course in public health statistics at Columbia University in 1999-2000 to demonstrate new technological data access methods, and a real time data acquisition and analysis was demonstrated using WebMIRS/SAS/SUDAAN at the CDC Data Users Conference in Bethesda, MD in July 2000.

WebMIRS enhancements done this year include packaging and deployment of WebMIRS as a Java application using Java WebStart technology, streamlining the capability to request image display, and redesigning Web pages for WebMIRS. In addition to the program enhancements, a new MySQL database was created to maintain information about WebMIRS usage, and an expert in Graphical User Interface (GUI) usability was consulted to analyze the WebMIRS interface. Also, collaborative work was initiated with Texas Tech University to develop a Wavelet compression capability to allow delivery of the WebMIRS images in compressed form. A software interface to support decompression and display of these images was designed.

WebMIRS version 1.0.9 currently has 150 users, both within and outside the U.S. Affiliations of recent users are: (i) New York Medical College, Valhalla, New York; (ii) Hospital de Apoyo La Merced, La Merced Chachamyo, Peru; (iii) Des Moines University Osteopathic Medical Center, Des Moines, Iowa; (iv) Hospital Materno de Santa Clara, Santa Clara, Cuba; and (v) TZI-University of Bremen, Bremen, Germany. The top three categories of use are currently General Research, Image

Interest, and Epidemiology.

Current WebMIRS work includes further development of the GUI to retrieve and display the radiological data, and incorporation of the recommendations of the WebMIRS GUI usability study.

The Digital Atlas of the Spine. This is a dataset of cervical spine and lumbar spine images with interpretations validated by a consensus of medical experts, along with software to display and manipulate the images. The images in the Atlas were chosen from the 17,000 images collected in the NHANES II survey. We convened two workshops in collaboration with other National Institutes of Health researchers to seek expert advice and consensus on a wide set of technical and biomedical issues related to the radiological interpretation of this set of images. Among the issues covered were the exact features to be interpreted. Radiographic features considered for interpretation of the cervical images were anterior osteophytes, posterior osteophytes, disc space narrowing, sclerosis, vacuum phenomenon, and subluxation. For the lumbar images, features considered included anterior osteophytes, posterior osteophytes, disc space narrowing, sclerosis, vacuum phenomenon, spondylolisthesis, spondylolysis, and DISH. A subset of these features was selected as likely to be consistently interpretable from the NHANES images. This selection of features, based on the consensus of experts at the workshop, took into account published studies relating to the likelihood of obtaining consistent readings for the features considered. The features identified by the workshop as consistently readable were those chosen for the Atlas.

For the cervical spine images, the Atlas contains numerical interpretations or "grades" for anterior osteophytes and disc space narrowing, on a scale from 0-3, with 0 being "normal" and 3 being "most abnormal"; and also interpretations for subluxation, on a 0-1 scale, with 0 being "normal" and 1 being "abnormal". Similarly, for the lumbar spine images, the Atlas contains interpretations for anterior osteophytes and disc space narrowing, on a scale from 0-3. The Atlas user may display single or multiple images in order to view, for example, all grades from *normal* to *most abnormal* of anterior osteophytes in the cervical spine. Image processing capability is provided to assist in contrast enhancement for viewing of detail.

The Atlas may be accessed either as a Java applet, or downloaded as a Java application, from the CEB website. In addition, we provide a version of the Java application on CD. The Java application version allows the user to add his/her own images (either grayscale or color) in a special "My Images" section, and to annotate and title those images for later use.

This year the Atlas was enhanced by the addition of flexible capabilities for user annotation of images using a comprehensive set of menu-selectable drawing tools.

Version 2.1 of the Atlas is now being distributed. Though started and led by research engineers at CEB, this project has the following collaborators: the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), the NIH Clinical Center, and rheumatology experts from the University of Miami, the University of California at San Francisco, and Johns Hopkins University.

Online x-ray archive. The complete set of 17,000 NHANES II x-ray images in the full-resolution form in which they were digitized were made publicly available in fiscal year 2000. These images are available by FTP access to the CEB main network server, and have been accessed by researchers from both within the U.S. and also from international sites. For viewing the x-rays, we have created the ImViewJ software, a Java application that may be downloaded from our Web site and which allows the viewing of the images at their full spatial resolutions (1463x1755 for the cervical spine images, 2048x2487 for the lumbar spine images. For 550 images we also have coordinate data collected under the supervision of a radiologist at Georgetown University. This coordinate data defines landmark points for each vertebra in a manner commonly used in the field of vertebral morphometry, and serves as reference data to aid in creating and evaluating the performance of image processing algorithms for segmentation of the vertebrae. This coordinate data is publicly available on the FTP site along with TIFF 8-bit versions of the corresponding x-ray images. Users may access this coordinate data either through the FTP archive or through the WebMIRS system.

In FY 2001, a database was created to track usage of the FTP archive. There are currently 160 users, of which the most recent are: (i) Queensland University of Technology, Brisbane, Australia; (ii) TZI-University of Bremen, Bremen, Germany; (iii) CCS SA, Athens, Greece; (iv) San Francisco Center of Oral and Facial Surgery, San Francisco, California; (v) King Mongut University of Technology, Bangmod, Thailand. The top three categories of use are currently Image Processing, Medical Research, and Medical Education.

Image and coordinate data from this archive was used extensively in the 2002 Ph.D. thesis published as Strings and Necklaces: On learning and browsing medical image segmentations, by S. Ghebreab, University of Amsterdam. This data is also being used in doctoral work at Texas Tech University and other institutions.

Content-Based Image Retrieval (CBIR). The overall goal of this research is to develop methods for effective extraction of biomedical information from digital images of the spine. This work has implications both for indexing of image data and for search of that data. For example, for the 17,000 NHANES II images, the only indexing data available is the collateral (alphanumeric) data collected in the questionnaires and examinations; no indexing information derived directly from the images is available, and the high cost of employing radiological experts to compile such data by physical viewing and interpreting each image makes it unlikely that such information will ever be acquired by purely manual means. These circumstances could be reversed if reliable, biomedically-validated software could produce image interpretations automatically. Even in the more likely case that only semi-automated methods should prove feasible, the reduction in labor costs could be sufficient to allow the creation of databases of significant biomedical information where this is not currently economically feasible. This is the implication of research into computer-assisted image indexing. Computer-assisted image searching is a potential enabler of enhanced information extraction from a database that has already been indexed. The most popularized form of this type of search is *query by example* or a variant, *query by sketch.* In query by example, the user inputs an image, perhaps by selecting from a set of choices provided by the system, or by providing a completely new image, and queries the database by asking, in effect, "Find records with images like this one", usually with respect to one or more characteristics of the example image, such as shape, histogram, or texture. In query by sketch, the input image is replaced by a sketch by the user, using drawing tools provided by the system. In either case, the system analyzes the input into component features, then searches the images in the database for those with similar features. Results are usually returned as a similarity ranking.

An initial prototype Content-Based Image Retrieval system (CBIR1) was implemented for the retrieval of images based on simple vertebral shape models. This MATLAB program allows the user to specify up to 9 control points and the geometric configuration of these points to define an approximate vertebral shape to search for. The prototype database contains 100 cervical and lumbar images, and will rotate and scale each vertebra in each image to identify the best match to the input shape. Alternatively, the user may specify an example vertebra, and the program will search for the best shape match to the example.

During the current year accomplishments in this area included (1) implementation of a new, more capable Content Based Image Retrieval prototype system (CBIR2); (2) the implementation of the Active Shape Modeling algorithm by our collaborators at Texas Tech University; and (3) work in classifying vertebrae for normal/abnormal anterior osteophytes and disc space narrowing by the use of artificial neural networks.

The second CBIR prototype (CBIR2) was created with significantly enhanced capabilities, including an *indexing function* with capability to do active contour segmentation, to create detailed representations of vertebrae boundaries, and to convert these boundaries into multiple shape representations (global shape descriptors, invariant moments, polygon turn functions, and Fourier descriptors). In addition, a *retrieval function* was implemented that supports retrieval of shapes by any of the above shape representations. The database created includes NHANES text data as well, and supports query by sketch, image example, and text, in addition to hybrid text and image-based queries. In order to create this system, research was conducted to select the best candidate shape representations. The MySQL database system was incorporated into the retrieval function for the storage and retrieval of the text data. Current CBIR work is directed toward continued completion of segmentation functions for indexing, analysis of effectiveness of the various shape methods implemented for spine x-rays with significant osteoarthritis features, implementation of spatial data trees for feature vector organization, and creation of a database of segmented vertebrae of significant size and segmentation accuracy, to serve as testbed data for the ongoing CBIR work.

An implementation of the standard Active Shape Modeling segmentation algorithm as developed by Dr. Tim Cootes of the University of Manchester, was created in collaboration with Texas Tech University, and work is underway to enhance the user interface to this implementation, and to extend the capability of this implementation for segmenting digital spine x-rays.

Feature classification work was conducted in collaboration with Dr. Joe Stanley of the University of Missiouri-Rolla and consisted of the implementation of artificial neural networks for classifying (as normal or abnormal) anterior osteophytes in the cervical and lumbar spines, and disc space narrowing in the cervical spine. “Truth” data for the training and performance evaluation of these networks was obtained from biomedical experts in the field of spine imaging. Results obtained were as follows: the classifier of anterior osteophytes in the cervical spine was trained on 352 vertebrae and tested on 352 different vertebrae, and 85% agreement with the “truth” standard was attained; the classifier of anterior osteophytes for the lumbar spine attained agreement of 71% (training set size was 391 vertebra; test set was 391 different vertebra). For disc space narrowing, 85% agreement was attained (training set was 50 images; test set was 50 images). Feature classification work is continuing with classification efforts directed toward classifying lumbar spine images for disc space narrowing, cervical spine images for subluxation, and lumbar spine images for spondylolisthesis.

Creation of an Image Processing Resources page on the CEB Web site. In 2002, a Web page was created to hold important materials related to CEB-sponsored image processing work. This page contains synopses of work done in collaboration with CEB in the areas of segmentation and feature classification and provides downloadable versions of presentations of significant accomplishments in these areas.

AnatQuest: A window into the Visible Human

Building on the AnatLine system, the object-oriented database of Visible Human images indexed for the male thorax region, we created AnatQuest with the goal of providing widespread access to the VH images for users with low speed connections as well. AnatQuest offers users thumbnails of the cross-section, sagittal and coronal images of the Visible Male, from which detailed (full-resolution) views are accessed. Low bandwidth connections are accommodated by a combination of adjustable viewing areas and image compression done on the fly as images are requested. Users may zoom and navigate through the images. Since release in June 2002, it has averaged about 60,000 hits per month, about 5 times the number of hits for AnatLine.

In addition to its main purpose, AnatQuest serves as an access point for AnatLine which allows access through anatomic terms to high resolution cross-sectional images and segment masks (useful for rendering anatomic objects). The tools needed to use AnatLine are also available: VHParser and VHDisplay. The first is for unpacking the data files into its individual components (cross-section images, byte masks, coordinate and label tables, etc.). VHDisplay is for displaying both cross-sectional and rendered images. Also, VHDisplay is augmented to voice names of anatomic structures as the images are displayed.

Other resources accessible through AnatQuest are: 195 surface-rendered objects created at the Lister Hill Center as well as from outside sources (e.g., VoxelMan); and the FTP server for bulk transfer of high resolution image files.

In addition to the web-mediated version of AnatQuest, a 'kiosk' version was developed for the Dream Anatomies exhibit at NLM as a Java application suitable for onsite patrons with operation through a touchscreen monitor.

Concurrent with developmental activities, research proceeded as follows: (a) An investigation was undertaken to evaluate the use of Java servlets for the viewer functions in place of the applets currently implemented; the servlet accesses the database and transfers the data to an HTML page which is posted to the client machine; the advantages are that Java plugins are not required at the client and since HTML is supported by all browsers, the differences between IE and Netscape are not relevant; disadvantages: limited user interaction and fewer options for esthetic design of the user interface. (b) Compression issues were investigated; for better compression, all 1,878 CCD image slices had blue background changed to black.

Turning The Pages Information Systems

In 2001-2002, NLM and the British Library collaborated in the production of two virtual books: Blackwell's Herbal and Vesalius' Anatomy in photorealistic "Turning The Pages" form. The pages of these were scanned, and these high quality color images were manually processed by Adobe Photoshop and animated by Macromedia Director software, and displayed on a touchscreen monitor. The library patron may 'touch and flip through' each of these books.

Using the TTP books as a starting point, we began an investigation of different ways to extend them beyond their application as beautiful museum pieces to information systems (TTP+). We followed two different approaches, the 'discovery' and the 'storyline' models, to implement the extensions.

The TTP+ version of Blackwell's Herbal (discovery model) retains the photorealism of the original TTP, while allowing a patron to 'travel' to live sites on the Internet. E.g., from highlighted text on the St. John's Wort page, one can go to a PubMed search and get citations, or go to ClinicalTrials.gov and get information on clinical trials of this drug. Also, links are available for plant descriptions and photographs on sites of the USDA, Forest Service and U.S. Herbaria, among others.

Problems to be resolved include the existence of inappropriate links (e.g., commercial sites) from some legitimate resources, and links that open a separate browser window. To address the first problem, a list was made of all URLs from the Blackwell plant pictures and resources, and reviewed for appropriateness. For the second problem, a perl script was created to automate the following steps: when a link opens a separate browser window, clone the page, and edit the HTML to make the link come up within the main browser window.

The TTP+ version of Vesalius followed the storyline model. Here, while elements of the design strategy for the Blackwell TTP+ were employed (e.g., menu button to serve as a table of contents, animated page flipping, timeouts and countdown warnings), the page images and images from other sources (e.g., rendered Visible Human images, pictures of Italian cities, etc.) were interlinked to present the patron with several multimedia 'stories,' e.g., "Man of Padua," "Modes of portraying anatomy."

In both these cases, the TTP+ versions, by incorporating explanatory and current information, extends the TTP books to services useful for the information-seeking user. Our work in developing Blackwell TTP+ was recorded in a videotape available from the Lister Hill Center.

Next Generation Internet: Infrastructure development and applications

Multilateral Initiative on Malaria in Africa. In our engineering role in this project we developed a statement of work for performance evaluation of the network linking malaria research sites in Africa, and the procurement was awarded in March 2002 to Infinite Global Infrastructures (IGI). Mr. Mike Gill, CEB, accompanied by IGI personnel, conducted a review of Redwing Satellite Solutions, the space segment and Internet access provider located near London. Following this visit, on June 10 he gave a presentation on NLM's communications work and the performance measurement task at the NLM-sponsored MIMCom Sysops workshop at the Wellcome Trust/KEMRI Research Centre located in Kilifi on the Kenya coast.

Experimental work. Following an analysis of data from the Trans-Pacific Demonstration of Visible Human completed a year ago, CEB staff co-authored a paper with Japanese participants from Sapporo Medical University. The paper was published in the March 2002 issue of the Space Communications Journal.

NGI meetings. Engineering staff represented NLM at the following meetings: Joint Engineering Team (JET) meetings at the National Science Foundation; Internet2 Health Sciences Working Group.

Maryland Governor's Task Force on High Speed Networks

In 2002, the Lister Hill Center continued to serve as a federal representative to the Maryland Governor's Task Force on High Speed Networks and the Engineering Advisory Group. The Task Force developed a comprehensive plan for bringing the state's network infrastructure in line with the needs of the 21st century. This plan, completed and presented to the legislature, contains recommendations to: a. combine existing state resources to maximize the state's return on investment; b. use existing state owned fiber where available; c. use current right-of-ways the state possesses to add additional fiber in underserved regions such as the Eastern Shore, Western and Southern Maryland; d. provide equity of access to all regions of the state, and support multiple segments of our society; e. promote collaboration among businesses, educational institutions, governmental bodies and research institutions; f. conduct a select number of high priority pilot projects in health care, business infrastructure development, and state government functions. A major contribution by the Lister Hill Center was made in the development of pilot projects in health care involving remote oncology treatment planning and remote intensive care support.

Proteus Project

This two-year project concluded in early 2002. A Medical Informatics Fellow undertook an investigation of system architecture for using medical knowledge in the form of *executable* distributed components to construct clinical protocols and thereby to represent the clinical process. The goal was to research the design of a system for medical decision making, data entry and data storage in a clinical setting, In this approach, called Proteus (PROTocols Editable by USers), clinical processes are represented by three types of "knowledge components": actions, processes and events. Each such "KC" has a mechanism to infer its own value and to determine the next action to be launched.

A Java-based proof-of-concept module (Protean) based on the Proteus architecture was built, and as a way to demonstrate its operation, a clinical protocol *Magnesium Sulfate therapy for severe pre-Eclampsia/Eclampsia* was created and demonstrated. This year the editing feature in Protean was improved by (a) creating new activity links between different KCs by just dragging from one KC's connection points to another, and (b) dragging a new KC from a panel graphically displaying different KCs onto the protocol that is loaded and visible on the main Protean screen.

To identify a suitable inferencing tool, a comparison was made of CTX, the criteria based inferencing tool developed at NLM, and Jess which is a rule-based expert system written in Java. Jess is a Java clone of the famous CLIPS of NASA, and it is easier to access, since Protean is also written in Java. Jess also offers a rich set of functionality, which may be exploited at a later stage. These reasons determined the choice of Jess as the inference tool for Proteus.

Also, to avoid the need for a user to create complex rules for every KC created, a "user as an inference tool" feature was incorporated into Protean. When a KC with the user as its designated inference tool, has to make an inference, a dialog box is displayed showing the preconditions on which the decision has to be based along with all the decision-options that are valid in that situation in a combo-box. The user may then select the most appropriate one. As an option, the user could be provided with other support information like web pages or multimedia information that may support the decision-making.

Another role for the user-inference tool is to serve as a fall back mechanism, i.e., if the automated inference mechanisms fail to make a valid inference for any reason, the inference making is passed on to the user. With this capability we have further verified our hypothesis of pluggable inference tools. We can easily switch from one type of inference tool to another by simply selecting the desired one from a combo list.

One benefit of using clinically meaningful entities – the clinical knowledge components – is that new uses, which depend on clinical semantics, can be incorporated with relatively little effort. To demonstrate this aspect of the Proteus approach, some *just-in-time* features were introduced into Protean. The Just In Time project at the Lister Hill Center aims to create generic questions that can be instantiated for specific clinical situations. A key requirement for such an approach is to discover what the clinician is engaged with at any moment while managing a patient. The Proteus approach lends itself to addressing this requirement. If the user selects any transaction KC, a window opens and shows in a tree structure all the possible questions pertaining to the situation represented by the KC, organized into different categories. When the user selects the questions the user is interested in, and clicks on the "answer" button, a browser is opened with PubMed responses to a query string representing the question.

To provide JIT functionality each KC was associated with a "category" – which represents the generic nature of the KC in clinical terms (e.g., "drug", "test", "clinical finding"). Also, a "term" is associated with the KC (e.g., the KC dealing with suspected breast lump is associated with term "Malignant Neoplasm of Breast"), which represents the core concern of the KC. This is preferably a MeSH term. To help the user select the appropriate MeSH term, the Protean editor can find the appropriate MeSH equivalent of a term typed in the field by querying the UMLS Knowledge Source Server. The user may also type in a broader term and get a more specific term by clicking on the "narrow" button. This opens another window that shows, as a tree, the broad term and all its sibling terms as branches, with more specific terms shown as leaves.

Engineering Laboratories

The R&D conducted by the Communications Engineering Branch rely on laboratories designed, equipped and maintained by the Branch, as well as content resources that support research.

Document Imaging Laboratory. This laboratory supports DocView, MARS and other research and design projects involving document imaging. Housed in this laboratory are advanced systems to electro-optically capture the digital images of documents, and subsystems to perform image enhancement, segmentation, compression, OCR and storage on high density magnetic and optical disk media. The laboratory also includes high-end Pentium-class workstations running under Windows 2000, all connected by 100 Mb/s Ethernet, for performing document image processing. Both inhouse developed and commercial systems are integrated and configured to serve as laboratory testbeds to support research into automated document delivery, document archiving, and techniques for image enhancement, manipulation, portrait vs. landscape mode detection, skew detection, segmentation, compression for high density storage and high speed transmission, omnifont text recognition, and related areas.

The laboratory also contains rack-mounted, networked processors running all recent versions of Windows-based operating systems to support the DocView, DocMorph and MyMorph projects. This provides an easily-configurable test platform for simulating a variety of potential user environments, including those with firewalls, for testing, modifying and improving software developed in these projects.

Document Image Analysis Test Facility. Designed, developed and maintained by the Communications Engineering Branch, this off-campus facility houses high-end Pentium workstations and servers that constitute the MARS production system. While routinely used to produce bibliographic citations for MEDLINE, this facility also serves as a laboratory for research into techniques for autozoning, autolabeling, autoreformatting, intelligent spellcheck and other key elements of MARS. Besides real time performance data, also collected and archived are large numbers of bitmapped document images, zoned images, labeled zones, and corresponding OCR output data. This collection serves as ground truth data for research in document image analysis and understanding.

Image Processing Laboratory. The Communications Engineering Branch Image Processing Lab is equipped with a variety of high end servers, workstations and storage devices connected by 100 Mb/s Ethernet. The laboratory supports the investigation of image processing techniques for both grayscale and color biomedical imagery at high resolution. In addition to computer and communications resources and image processing equipment to capture, process, transmit and display such high-resolution digital images, the laboratory also has a variety of image content.

The equipment includes a Sun Ultra Enterprise 4000 server with dual 168 MHz CPUs, 512 MB memory, 18 GB of disk storage for application development, and two locally attached Sun StorEdge A5000 RAIDs. Additional computers in the lab include two Sun Ultra 10 workstations, each with a 440 MHz CPU, 512 MB memory and an external 18 GB SCSI disk, two Sun Ultra 10s each with a 330 MHz CPU and 512 MB memory, a Sun SPARCstation 10 with two 40 MHz CPUs and 256 MB memory, and a Sun SPARCstation 20 with two 50 MHz CPUs and 256 MB memory. All of these machines run the Sun Solaris 2.7 operating system.

Large-scale magnetic storage is provided by two RAID systems, the larger of which consists of a pair of Sun StorEdge A5000 storage arrays which are attached to the host via dual-looped fibrechannel connections with a maximum throughput of 200 MB/sec, and which provide approximately 150 GB of storage. A small Sun Sparc Storage Array Model 100 RAID system provides an additional 25 GB of storage.

Two ultra-high-resolution E-systems Megascan displays provide image display at spatial resolution of 2048x2560 pixels. An IBM-compatible PC and monitor are also available in the lab for PC testing of Java applications.

Most machines are equipped with multiple networking ports (FDDI, ATM, Ethernet, fast Ethernet) which allow, in addition to standard networking capabilities on the local Ethernet, the capability of alternate physical communications channels with these machines. This capability has been used in communications engineering experiments for point-to-point satellite channels connecting these machines with remote sites. ATM switches connect the Ethernet and FDDI networks to other local area networks throughout the building, to the Internet, and to experimental ATM, in addition to Abilene, the infrastructure for the Next Generation Internet and Internet-2 initiatives.

Image Processing Lab content resources. A large part of the NHANES II data has been put into the WebMIRS database tables. All of the NHANES II demographic, anthropometric, physical examination, and adult health questionnaire data is available through WebMIRS, as well as the statistical weighting and sampling strata variables required for analysis of the data. This data covers a sample of approximately 20,000 survey participants. In addition, the 17,000 NHANES II cervical and lumbar spine x-ray images are available for viewing through WebMIRS, in one-quarter spatial resolution format. These 17,000 images are stored in a magnetic RAID system and are available for public downloading via FTP.

Similarly, a large part of the NHANES III data has been put into the WebMIRS database tables. All of the NHANES III demographic, physical examination, health questionnaire, and laboratory data is available through WebMIRS, as well as the statistical weighting and sampling strata variables required for analysis of this data. The NHANES III data covers a sample of approximately 30,000 survey participants.

In addition to the data above, The Image Processing Lab also contains a selection of History of Medicine color images digitized at high resolution from the Library's Arabic and Persian medical manuscript collection.

CEB Publications for calendar year 2002

Thoma GR, Ford G, Le DX, Li Z. Text verification in an automated system for the extraction of bibliographic data. Proc. 5th International Workshop on Document Analysis Systems, Springer-Verlag: Berlin, August 2002, 423-32.

Le DX, Straughan SR, Thoma GR. Greek alphabet recognition technique for biomedical documents. Proc. 6th World Multiconference on Systemics, Cybernetics and Informatics, Vol. III, eds: Callaos N, et al; July 2002, 86-91.

Walker FL, Thoma GR. A SOAP-enabled system for an online library service. Proc. InfoToday 2002, Medford NJ: Information Today, May 2002, 320-9.

Tezmol A, Sari-Sarraf H, Mitra S, Long R, Gururajan A. Customized Hough Transform for robust

segmentation of cervical vertebrae from x-ray images. Proc. 5th IEEE Southwest Symposium on Image Analysis and Interpretation, Los Alamitos CA: IEEE Computer Society, April 2002, 224-28.

Nishinaga N, Tatsumi H, Gill M, Akashib A, Nogawa H, Reategui I. Trans-Pacific demonstration of Visible Human (TPD-VH). Space Communications 17:4, March 2002, 303-11.

Long LR, Krainak DM, Thoma GR. Identifying image structures for content-based retrieval of digitized spine x-rays. Proc. SPIE Medical Imaging 2002: Image Processing, Vol. 4684, February 2002, 1204-14.

Krainak DM, Long LR, Thoma GR. A method of content-based retrieval for a spinal x-ray image database. Proc. SPIE Medical Imaging 2002: PACS and Integrated Medical Systems, Vol. 4685, February 2002, 108-16.

Zamora G, Sari-Sarraf H, Mitra S, Long R. Analysis of the feasibility of using Active Shape Models for segmentation of gray scale images. Proceedings of SPIE Medical Imaging 2002: Image Processing.Vol. 4684, San Diego CA, February 2002, 1370-81.

Thoma GR, Ford G. Automated data entry system: performance issues. Proc. SPIE Document Recognition and Retrieval IX, Vol. 4670, January 2002, 181-90.